欠損率の高い交通行動データに含まれる無回答バイアスの修正方法

㈱福山コンサルタント　正会員　原田慎也
広島大学大学院国際協力研究科　正会員　藤原章正
広島大学大学院国際協力研究科　正会員　杉惠頼寧

１.はじめに

交通需要予測の際の基礎データを得るためにアンケート方式の調査を実施することが一般的である.しかしアンケート調査は回答者の主観データを得る調査であるため,データには必然的に無回答に伴う欠損データが存在し,その結果バイアスが生じる危険性がある.この問題の対策として,筆者らは統計的手法であるEMアルゴリズム[1)]を用いて無回答バイアスの修正に関する研究を行ってきたが,複数変数欠損を伴う高欠損率データに対しては,十分な修正効果を得ることができなかった(表１)[2)]. ここで無回答バイアスとは回答が欠損することによって観測される交通行動や行動と要因との因果関係が歪むことをいう.

一方,近年交通計画分野においても GIS の普及が著しく,各種デジタル情報の開発整備が進みつつある.このGIS を用いることによって,一定の範囲内であるが個人の移動毎に交通サービス水準に関する客観値を得ることが可能となる[3)].

そこで本研究では,アンケート主観データの欠損部分をGIS から求まる客観値で補填する.その際に統計的手法であるEMアルゴリズムを適用し,高欠損率データに含まれる無回答バイアスを除去することを目的とする.

２. 高欠損率データにGIS を用いた場合の修正効果の分析

(1)分析データの作成

基本データは,本研究室が1994 年の 11 月と 1997 年の 11 月に行った第１回,第２回新交通システム(通称:アストラムライン)に関する事後 RP 調査のデータである.

このデータを基に,一部データを欠損させた仮想データを作成した.アンケート回答値であるので以下,主観値と呼ぶ.本分析は新交通システムと自動車に着目して行うため,主要交通手段が新交通システムで代替交通手段が自動車,もしくは主要交通手段が自動車で代替交通手段が新交通システムのサンプルのみを抽出する.そのサンプルの中から,自動車の乗車時間,新交通システムの乗車時間,新交通システムへのアクセス時間の計３項目全てに答えているサンプルを抽出した.これを完全データと呼ぶ.

次にGISを用いた最短経路解析によって,自動車の乗車時間,新交通システムの乗車時間,新交通システムへのアクセス時間それぞれの客観値を算出した.ネットワークは本研究室が構築した広島都市圏ネットワークを使用した.

以上のように作成されたデータの概要を,表１に示す.

表１　データの各状況別における無回答バイアスの対処方法

		変数間相関：大	変数間相関：弱
欠損変数の数：少	欠損率：小	従来の方法	従来の方法
	欠損率：大	EM (改善効果：大)	EM (改善効果：小)
欠損変数の数：多	欠損率：小	EM (改善効果：大～小)	?
	欠損率：大	?	?

表２　分析データの概要

変数		平均値	標準偏差
車乗車時間(分)	主観値	43.69	20.24
	客観値	32.94	9.84
新交通乗車時間(分)	主観値	18.56	7.64
	客観値	16.81	7.09
新交通ｱｸｾｽ時間(分)	主観値	9.15	5.13
	客観値	13.66	5.79
機関選択結果(自動車:0,新交通:1)		0.32	0.47
サンプル数		96	

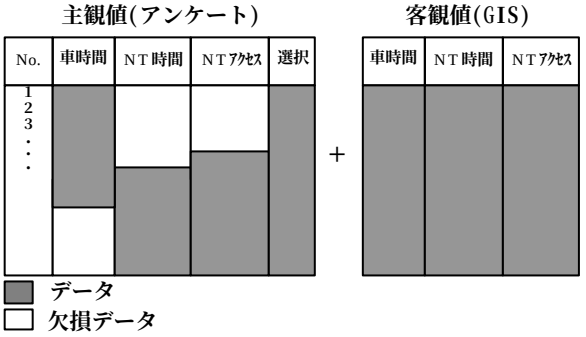

図１　分析の概念図

(2)分析方法

図１を用いて分析方法を簡単に説明する.欠損の存在する主観値データの４変数に３変数の客観値データを加え,計７変数の１つのデータセットと考えEMアルゴリズムによる修正を行う.主観値と変数間相関の強い客観値を加えることによって,主観値のみを用いた修正に比べて補完精度が格段に格段に改善されることが期待できる.

キーワード：調査論, 無回答バイアス, EMアルゴリズム, GIS
連絡先：739-8529　東広島市鏡山 1-5-1　TEL&FAX 0824-24-6921

(3)分析結果

①交通サービス変数の分布母数に含まれるバイアスの修正結果

主要交通手段が自動車,すなわち代替交通手段である新交通システムに関する主観値データ(新交通乗車時間と新交通アクセス時間)を欠損させた.

図２に欠損率を10%～80%の範囲で８段階に変えながら修正前のバイアスを含む平均値,主観値データのみを用いてＥＭアルゴリズムにより修正した後の平均値,主観値データに客観値を加えたデータを用いてＥＭアルゴリズムにより修正した後の平均値を比較した結果を示す.

修正結果(主観値のみ)と修正結果(主観値+客観値)を比較した場合,後者の方が修正効果が大きいことが明らかである.これは修正に用いることのできる変数が４から７に増えることによる.特に,欠損率が大きくなるにつれて改善率の差が大きくなることから,客観値を加えることによる修正効果が高欠損率データにおいて大きいことを確認することができた.

標準偏差の分析においても同様の修正効果を得た.

②モデルパラメータに含まれるバイアスの修正結果

分析に用いるモデルは２項選択非集計ロジット型であり,式(1)に示すような２つの変数からなる線形効用関数Vをもつものとする.なお,nは個人,iは選択肢を表す.

$$V_{in} = \beta_t \, time_{in} + \beta_a \, access_{in} \qquad (1)$$

図３は,上述の欠損パターンでの欠損率とβ_tの修正効果の推移を示したものである.図３を見ると,欠損値の補完を行わない欠損データのパラメータ値が欠損率の違いによって大きく変動しているのが分かる.各欠損率に応じて需要が過大にも過小にも評価されるようでは実用的でない.

一方,主観値データのみを用いたＥＭアルゴリズムによる修正効果と,それに客観値を加えたデータを用いたＥＭアルゴリズムによる修正効果を比較すると,どちらも過小評価ではあるが,後者の方が前者に比べて格段に修正されているのが分かる.その修正効果は高欠損率になるほど大きくなっている.

次に図４にパラメータβ_tの標準誤差の修正効果の推移を示す.図を見ると分かるように,欠損率に乗じて欠損データにおけるβ_tの標準誤差は大きくなる.ＥＭアルゴリズムを用いて修正を施すことによって,完全データに近いパラメータ推定値の有効性を得ることができることが明らかとなった.これに対して主観値データのみを用いた場合とそれに客観値を加えたデータを用いた場合の修正効果に大きな差を見ることはできなかった.

３．おわりに

本研究では高欠損率データにおける無回答バイアスの修正方法を提案した.分析結果より,高欠損率データにGISから求まる客観値を加えＥＭアルゴリズムを用いることによって,データの分布母数及びモデルパラメータの普遍性を高めることができることが明らかになった.

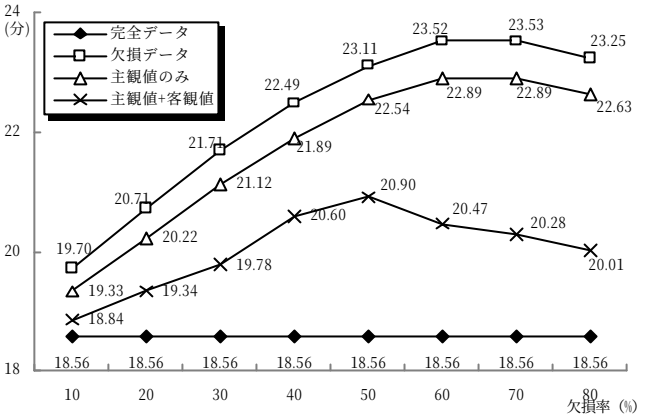

図２　欠損率と新交通乗車時間(平均値)の推移

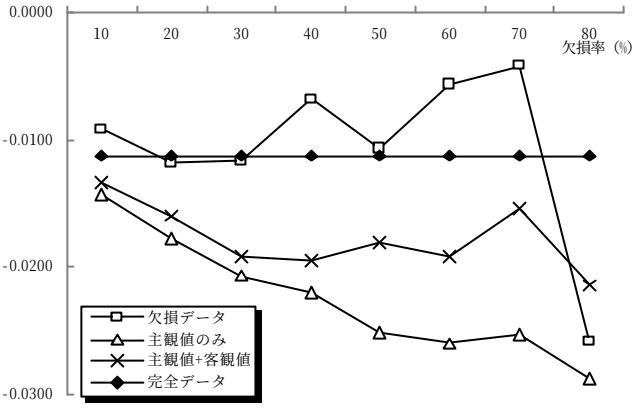

図３　欠損率とパラメータの修正効果の推移

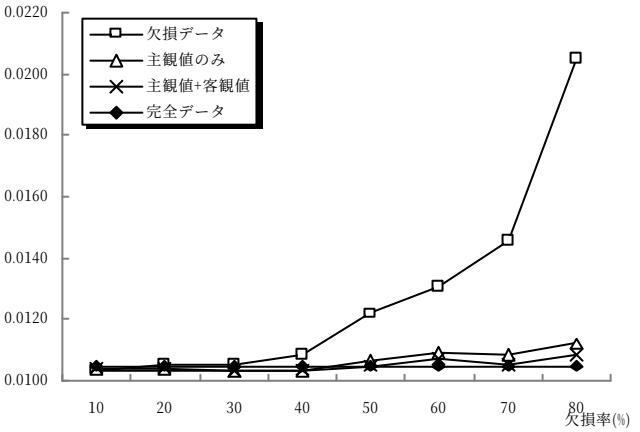

図４　パラメータの標準誤差の推移

参考文献

1) R. Little and D. Rubin : Statistical Analysis with Missing Data, John Willy and Sons, 1987

2) 藤原,杉惠,原田：交通日誌データにおける無回答バイアスの修正方法，土木計画学研究･論文集，No.16, pp121-128, 1999.

3) 藤原,杉惠,山下：GISを活用した交通日誌の欠損データの補填方法，土木計画学研究･講演集, No.22 (2), pp407-410, 1999.